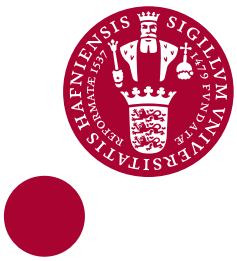

Comparative genomics for non-O1/O139 Vibrio cholerae isolates recovered from the Yangtze River Estuary versus V. cholerae representative isolates from serogroup O1

Gong, Li; Yu, Pan; Zheng, Huajun; Gu, Wenyi; He, Wei; Tang, Yadong; Wang, Yaping; Dong, Yue; Peng, Xu; She, Qunxin; Xie, Lu; Chen, Lanming

Published in:
Molecular Genetics and Genomics

DOI:
10.1007/s00438-018-1514-6

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Molecular Genetics and Genomics (2019) 294:417–430
https://doi.org/10.1007/s00438-018-1514-6

ORIGINAL ARTICLE

Comparative genomics for non-O1/O139 *Vibrio cholerae* isolates recovered from the Yangtze River Estuary versus *V. cholerae* representative isolates from serogroup O1

Li Gong[1] · Pan Yu[1] · Huajun Zheng[2] · Wenyi Gu[2] · Wei He[3] · Yadong Tang[1] · Yaping Wang[1] · Yue Dong[4] · Xu Peng[5] · Qunxin She[5] · Lu Xie[6] · Lanming Chen[1]

Received: 22 August 2017 / Accepted: 13 November 2018 / Published online: 28 November 2018

Abstract

Vibrio cholerae, which is autochthonous to estuaries worldwide, can cause human cholera that is still pandemic in developing countries. A number of *V. cholerae* isolates of clinical and environmental origin worldwide have been subjected to genome sequencing to address their phylogenesis and bacterial pathogenesis, however, little genome information is available for *V. cholerae* isolates derived from estuaries, particularly in China. In this study, we determined the complete genome sequence of *V. cholerae* CHN108B (non-O1/O139 serogroup) isolated from the Yangtze River Estuary, China and performed comparative genome analysis between CHN108B and other eight representative *V. cholerae* isolates. The 4,168,545-bp *V. cholerae* CHN108B genome (47.2% G+C) consists of two circular chromosomes with 3,691 predicted protein-encoding genes. It has 110 strain-specific genes, the highest number among the eight representative *V. cholerae* whole genomes from serogroup O1: there are seven clinical isolates linked to cholera pandemics (1937–2010) and one environmental isolate from Brazil. Various mobile genetic elements (such as insertion sequences, prophages, integrative and conjugative elements, and superintegrons) were identified in the nine *V. cholerae* genomes of clinical and environmental origin, indicating that the bacterium undergoes extensive genetic recombination via lateral gene transfer. Comparative genomics also revealed different virulence and antimicrobial resistance gene patterns among the *V. cholerae* isolates, suggesting some potential virulence factors and the rising development of resistance among pathogenic *V. cholerae*. Additionally, draft genome sequences of multiple *V. cholerae* isolates recovered from the Yangtze River Estuary were also determined, and comparative genomics revealed many genes involved in specific metabolism pathways, which are likely shaped by the unique estuary environment. These results provide additional evidence of *V. cholerae* genome plasticity and will facilitate better understanding of the genome evolution and pathogenesis of this severe water-borne pathogen worldwide.

Keywords *Vibrio cholerae* · Comparative genomics · Mobile genetic elements · Virulence · Antimicrobial resistance · Estuary

Introduction

Vibrio cholerae is a Gram-negative bacterium that is autochthonous to estuaries worldwide (Kaper et al. 1995). The bacterium is the aetiological agent of cholera, a life-threatening human diarrhoeal disease (Colwell 1996; Heidelberg et al. 2000). Cholera has been epidemic in southern Asia for at least 1000 years, but has also spread worldwide to cause seven pandemics since 1817 (Wachsmuth et al. 1994). The seventh pandemic that erupted in 1961 has lasted for over 50 years (Feng et al. 2008). In recent years, cholera remains endemic in developing countries where sanitation is poor and drinking water unsafe (Heidelberg et al. 2000; Zhang

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-018-1514-6) contains supplementary material, which is available to authorized users.

✉ Lu Xie
xielu@scbit.org

✉ Lanming Chen
lmchen@shou.edu.cn

Extended author information available on the last page of the article

and Gou 2014). For example, the most recent outbreak was reported in Yemen on 27 April 2017, and the disease caused a total of 332,658 suspected cholera cases and 1,759 deaths till 13 July 2017 (World Health Organization, http://www.who.int/). However, the epidemiology and pathogenesis of *V. cholerae* are intricate; therefore, there is a clear need for a global genome-level understanding of the bacterium, particularly of its environmental origin.

Vibrio cholerae isolates have been classified into at least 206 serogroups (O1 to O206) based on their outer membrane O antigens, among which only serogroups O1 and O139 can cause cholera epidemics (Kaper et al. 1995). The serogroup O1 is further classified into two biotypes: classical and EL Tor, resulted in the sixth and the seventh cholera pandemics, respectively (Wachsmuth et al. 1994; Faruque and Mekalanos 2012). The crucial virulence determinants in pathogenic isolates are a cholera toxin (CT) related to a temperate filamentous phage CTXΦ (Heidelberg et al. 2000) and the receptor toxin co-regulated pilus (TCP) for entry of CTXΦ into the cell (Faruque and Mekalanos 2012). Previous studies have indicated that non-O1/O139 serogroups that embody genetically diverse strains also result in sporadic disease in a CT- and TCP-independent manner using undefined virulence mechanisms (Chatterjee et al. 2009). For example, the clinical isolate *V. cholerae* AM-19226 that does not produce CT or TCP still causes a rapidly fatal diarrhoeal disease due to transferred genes that are responsible for the type III secretion system (Shin et al. 2011). *Vibrio cholerae* LMA3984-4 ($ct^{-}tcp^{-}$) that was isolated from superficial water from the Tucunduba Stream in Brazil in 2007 displayed genetic similarity with epidemic strains with virulence-related genes (Sá et al. 2012). Acquisition of virulence or resistance traits via lateral gene transfer (LGT) might occur at high frequency throughout microbial communities in environmental ecosystems (Thompson et al. 2004). Mobile genetic elements (MGEs) that mediate LGT between bacteria, such as insertion sequences (ISs), prophages, integrative and conjugative elements (ICEs), and super intergons (SIs) (Wozniak et al. 2009), could constitute important driving forces in genome evolution and speciation of *Vibrios* (Thompson et al. 2004). Evidently, these mobile gene pools have rapidly crossed species boundaries to create endemicity within global *V. cholerae* populations (Boucher et al. 2011). For example, several waves of global transmission in the seventh cholera pandemic were observed, among which waves 2 and 3 were characterised by *V. cholerae* strains with the acquisition of SXT/R391 family ICEs (Waldor et al. 1996; Mutreja et al. 2011).

In 2000, the complete genome sequence of *V. cholerae* El Tor N16961, which was isolated in Bangladesh in 1971 and linked to the seventh cholera pandemic, was determined (Heidelberg et al. 2000). Along with the technological breakthroughs in genome sequencing technologies (Metzker 2005), the number of genome sequencing projects focused on the other *V. cholerae* isolates has distinctly increased. To date, more than 30 complete and numerous draft genome sequences for *V. cholerae* strains are available in the GenBank database (http://www.ncbi.nlm.nih.gov/genome/) or online (http://www.genomesonline.org). Many comparative analyses of these *V. cholerae* genomes have been reported (e.g., Banerjee et al. 2014; Dutilh et al. 2014; Okada et al. 2014; Garrine et al. 2017; Imamura et al. 2017). In the present study, we determined the complete genome sequence of *V. cholerae* CHN108B (non-O1/O139 serogroup, $ct^{-}tcp^{-}$), which was recently isolated from surface water from the Yangtze River Estuary, China (Song et al. 2013), to address the lack of complete genome data regarding this bacterium originating from an estuary environment in China (Yi et al. 2014). We also conducted comparative genomic analysis between CHN108B and the 8 representative *V. cholerae* isolates available online when the project began, including N16961, M66-2, O395 uid58425, and O395 uid159869, IEC224, MJ-1236, LMA3984-4 and 2010 EL-1786 (Supplementary Table S1). Among these serogroup O1 isolates, only LMA3984-4 originated from the environment, whereas the others were clinical, as follows: M66-2 is the 1937 Makassar outbreak isolate (Feng et al. 2008); O395 is linked to the sixth cholera pandemic (Feng et al. 2008; Chun et al. 2009); and the remaining isolates are linked to the seventh cholera pandemic (Heidelberg et al. 2000; Chun et al. 2009; Reimer et al. 2011; Garza et al. 2012). The data from this study have refined our grasp on the genome evolution and pathogenesis of the common water-borne pathogen worldwide.

Materials and methods

Vibrio cholerae strains and genomic DNA preparation

Vibrio cholerae CHN108B was isolated from surface water from the Yangtze River Estuary in Shanghai, China in 2011 (Song et al. 2013). The bacterium was identified as non-O1/O139 serogroup and non-toxic ($ct^{-}tcp^{-}$) in a previous study. *Vibrio cholerae* CHN108B was inoculated from our laboratory storage at − 80 °C into 5 ml Luria–Bertani (LB) broth (3% NaCl, pH 8.5) (Beijing Land Bridge technology Co. Ltd., China) and incubated at 37 °C with shaking. The overnight bacterial culture was then inoculated (1:100, v/v) into 200 ml LB broth. Bacterial cells grown to the logarithmic growth phase at 37 °C were immediately harvested by centrifugation at 2700*g* for 10 min at 4 °C. Genomic DNA was prepared using the Biospin Bacteria DNA Extraction Kit (BIOER Technology, Hangzhou, China) and plasmid DNA was isolated using the TaKaRa MiniBEST Plasmid Purification Kit Version 3.0 (Japan TaKaRa BIO, Dalian Company,

China) according to the manufacturers' instructions. Three independently prepared DNA samples were examined by agarose gel electrophoresis, visualised and imaged as described previously (He et al. 2015b). No plasmid DNA was extracted from CHN108B. Only pure genomic DNA samples (a 260/280 nm absorbance ratio of 1.8-2.0) were used for genome sequencing. The concentrations of the DNA samples were determined using a BioTek Synergy™ 2 multi-mode microplate reader (BioTek Instruments, Inc., VT, USA). Multiple *V. cholerae* isolates (CHN001f to 009f) recovered from the Yangtze River Estuary were identified as non-O1/O139 serogroup and non-toxic ($ct^{-}tcp^{-}$) and their genomic DNA were prepared as well.

Genome sequencing, assembly and gap closure

Whole-genome sequencing of *V. cholerae* CHN108B was conducted at the Chinese National Human Genome Centre (Shanghai, China) using the Genome Sequencer FLX (GS-FLX) Titanium system (Roche, Roche Applied Science, Basel, Switzerland). The obtained sequencing reads were assembled using the Roche/454 Newbler v2.3 software (https://lifescience.roche.com/), and the generated contigs (> 2000 bp) served as the basis for genome gap closure. The relationship between each large contig was determined by genome sequence alignment with reference genomes using the MUMmer 3.2.3 software (http://www.tigr.org/software/mummer/) (Kurtz et al. 2004). The chosen references with higher Megablast bitscore hits were complete genome sequences from *V. cholerae* N16961 and IEC224 (Supplementary Table S1). The Primer 5.0 software (http://www.premierbiosoft.com) was used to design PCR primers for genome walking as described previously (Chen et al. 2005). Long-range PCR amplification was performed using the Takara LA Taq kit (Japan TaKaRa BIO, Dalian Company, China) according to the manufacturer's instructions. All PCR amplifications were performed in a Mastercycler® pro PCR thermal cycler (Eppendorf, Hamburg, Germany). PCR products were analysed by agarose gel electrophoresis, purified using the AxyPrep DNA Gel Extraction Kit (Axygen, Silicon Valley, USA), and then sequenced using an ABI PRISM 3730XL DNA Sequencer (Applied Biosystems, Inc, Carlsbad, California, USA) for genome gap closure. Low-value sequences were confirmed by PCR and sequencing analysis.

Genome annotation

Open reading frames (ORFs) were predicted using the GLIMMER 3.02 software (Delcher et al. 1999). Functional assignments were inferred based on standalone Basic Local Alignment Search Tool (BLAST) (http://www.ncbi.nlm.nih.gov/BLAST) searches against an non-redundant protein database from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) and the Clusters of Orthologous Groups (COG) database (Tatusov et al. 2001). Each predicted protein was annotated if it met the criteria of a minimum cutoff of 30% identity at the amino acid level. Each gene was also functionally classified by assigning a COG number. The ORFs that had no hits against the non-redundant protein database (NCBI) were annotated as hypothetical proteins. The rRNA genes were annotated using the FgenesB tool (http://softberry.com), and tRNA genes were detected using the tRNAscan 1.21 software (Lowe and Eddy 1997). Putative virulence factor and antimicrobial resistance genes were detected using the Virulence Factors Data Base (http://www.mgc.ac.cn/VFs) and the Antibiotic Resistance Genes Database (http://ardb.cbcb.umd.edu/), respectively.

Comparative genome analysis

Comparison of whole chromosome sequences was performed using Mauve version 20150226 build 10 (c) package (http://darlinglab.org/mauve) (Darling et al. 2004) and the MUMmer 3.2.3 software. The Blastcluster software (http://www.ncbi.nlm.nih.gov/) was used for pan-genome analysis. Orthologous proteins were assigned only for proteins sharing both 60% amino acid identity and 80% sequence coverage, while those with lower than 30% or no hits were assigned as strain-specific genes. All the programmes were performed using the default parameters. The complete genome sequences of the eight representative *V. cholerae* isolates were retrieved from the GenBank database, each of which contains two chromosomes (chrs), including M66-2 (Feng et al. 2008), O395 uid159869 (Feng et al. 2008), O395 uid58425 (Chun et al. 2009), N16961 (Heidelberg et al. 2000), IEC224 (Garza et al. 2012), MJ-1236 (Chun et al. 2009), 2010 EL-1786 (Reimer et al. 2011), and LMA3984-4 (Sá et al. 2012) (Supplementary Table S1).

Analysis of mobile genetic elements

Insertion sequences (ISs) were identified using the ISFinder (Siguier et al. 2006). Prophages were identified using the Prophage Finder (http://phast.wishartlab.com). Super-integrons (SIs) were identified as follows: (1) *Vibrio cholerae* repeats (VCRs), such as TAACAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCCCATTGAGCCGCGGTGGTTTCGGTTGTTGTGTTTGAGTTTAGTGGTAATGCGTTGCCAGCCCCTTAGGCGGGCGTTAT (Mazel et al. 1998), were searched in the *V. cholerae* genomes by BLAST analysis; (2) identified VCRs were analysed based on a specific RYYTAAC+at least one ORF+GTTARRY structure (Rowe-Magnus et al. 2003); and (3) the number of SI cassettes was identified according to the term in the

cassette that is composed of two VCRs and at least ORF in the middle. The SXT/R391 family of integrative conjugative elements (ICEs) were identified as follows: (1) the ICEs-chromosomal junction sequences from the *prfC* gene, which encode a non-essential peptide release factor 3 in *Escherichia coli, V. cholerae* and other hosts, were searched in the *V. cholerae* genomes by BLAST analysis and (2) the sequences within the predicted *prfC* sequences were compared with SXT/R391 ICEs to identify conserved core genes in their modules, hotspots, and variable regions.

Results

General genome features of *V. cholerae* CHN108B

The whole genome sequence of *V. cholerae* CHN108B (non O1/O139 serogroup) was determined using the 454-pyrosequencing technique, which yielded 339,285 reads with a genome sequencing depth of 33-fold. These sequencing reads were assembled into 90 contigs (> 2 kb). We designed 450 oligonucleotide primers for genome gap closure, including 416 for closing sequencing gaps by genome walking and 34 for physical gaps by combinatorial PCR. The *V. cholerae* CHN108B genome consisted of two circular chromosomes that contained 3,083,301 bp (chr 1) and 1,085,244 bp (chr 2) (Fig. 1, Supplementary Table S2). The complete genome of the bacterium contained 4,168,545 bp with an average G+C content of 47.2%. The CHN108B genome was the second largest among the nine *V. cholerae* genomes analysed in this study; it contained 538 more protein-encoding genes than the environmental strain *V. cholerae* LMA3984-4 that had the smallest genome size (3,738,718 bp) (Sá et al. 2012). In the CHN108B genome, the 3691 predicted protein-encoding genes were grouped into twenty-two gene functional catalogues that were identified in the COGs database (data not shown), of which approximately 24% encoded hypothetical proteins with currently unknown functions in the public databases.

Consistent with the other *V. cholerae* genomes analysed in this study, the CHN108B genome contained recognisable genes for essential cell functions, such as DNA replication, transcription, translation and cell-wall biosynthesis. Most of the genes that encode enzymes for the predicted central metabolic pathways were also present in CHN108B, including those required for glycolysis, oxidative phosphorylation and the tricarboxylic acid cycle (TCA). The genes responsible for DNA restriction and modification and DNA repair (such as base excision repair, nucleotide excision repair and mismatch repair) were identified in CHN108B. Additionally, the

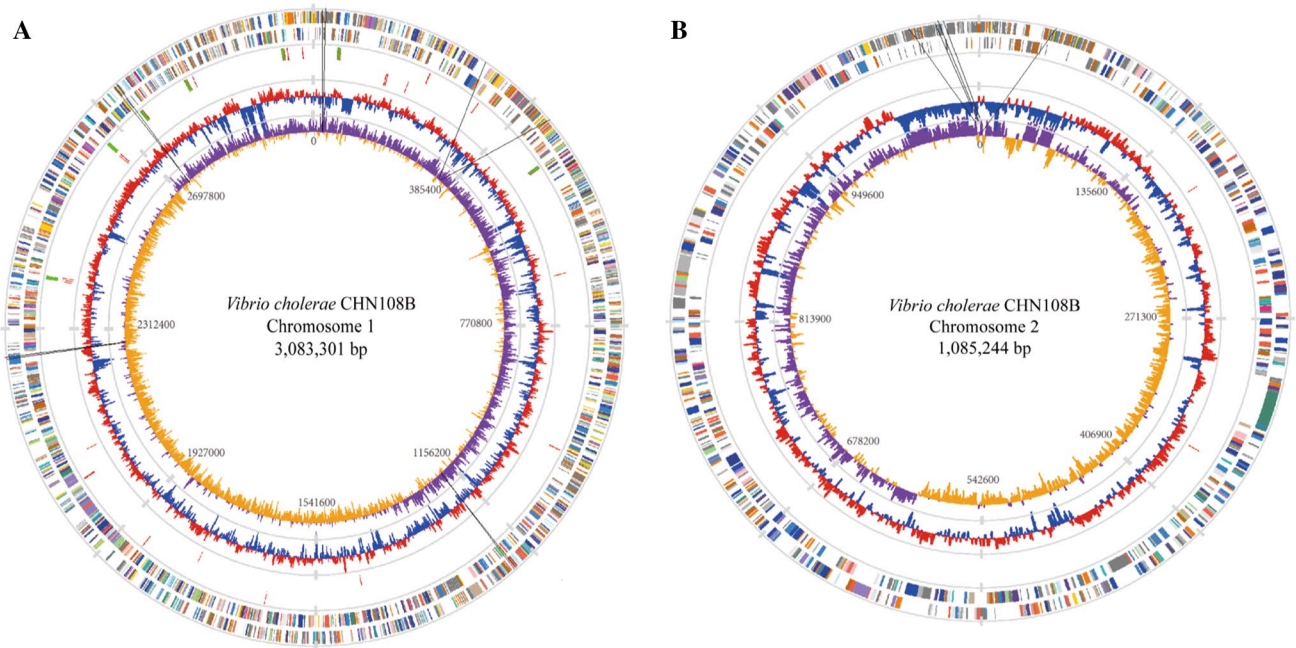

Fig. 1 Circular maps of *V. cholerae* CHN108B chromosomes. **a, b** Represent the larger and smaller chromosomes of *V. cholerae* CHN108B, respectively. Each circle in grey lines except for the two innermost circles illustrates different features on the plus (outer region) and minus (inner region) strands. Lines and boxes in the two outermost circles are coloured according to the COG categories. From the outside to inwards: the first circle, predicted protein-coding genes; the second circle, tRNA genes (in red) and rRNA operons (in green); the third circle, GC content (values higher than the average in red, and lower than the average in blue); the fourth circle, GC-skew (values more than zero in purple, and less than zero in orange). Relative positions of the MGEs were indicated in grey lines across the circles

CHN108B genome also carried numerous transposase genes (27) and various MGEs including ISs, ICEs, prophages and SIs, indicating the potential for LGT of hosted genes.

In marked contrast to the other *V. cholerae* genomes analysed in this study, the CHN108B genome lacked the *tupABC* genes required for the tungstate transporter system, including an ATP-binding cassette (ABC) transporter substrate-binding protein (*tupA*), an ABC transporter permease (*tupB*) and an ABC transporter ATP-binding protein (*tupC*), which indicated that CHN108B may not be able to utilise tungstate. CHN108B also lacked the *ugpC* gene (encodes a glycerol-3-phosphate transporter ATP-binding subunit) required for the sn-Glycrol 3-phosphate transporter system (*ugpBAEC*), suggesting an inactive monosaccharide transporter, similar to LMA3984-4. The comparative genomics also revealed that the CHN108B genome had 110 strain-specific genes, the highest number among the *V. cholerae* genomes analysed in this study, the majority of which encoded hypothetical proteins, while the rest were involved in secondary metabolism, and cell envelope and outer membrane biogenesis. For example, the CHN108B genome possesses the *rfbABC* gene cluster which shows high homology with the streptomycin biosynthetic gene cluster, including a glucose-1-phosphate thymidylyltransferase (*rfbA, chr1_02614*), a dTDP-glucose 4,6-dehydratase (*rfbB, chr1_02613*) and a dTDP-4-dehydrorhamnose 3,5-epimerase (*rfbC, chr1_02615*). CHN108B also had a gene encoding a limonene 1,2-monooxygenase (*chr1_01246*) for limonene and pinene degradation, which could be highly beneficial to the bacterium to utilise these compounds for increasing fitness in variable environments. The gene (*chr1_02000*) encoding a homologue of *Escherichia coli* curli production assembly/transport component CsgG was also identified to be specific to CHN108B. It will be interesting to determine its precise function in biofilms and other community behaviours in *V. cholerae*.

Genome structure

The whole genome sequences of the nine *V. cholerae* isolates were compared at a global level using the Mauve version 20150226 software. The resulting data are illustrated in Fig. 2. Six local collinear blocks (LCBs) were identified in the chr 1 s and shared by all the isolates (Fig. 2a); only two LCBs (24.1 and 151 kb) showed a change in relative genomic position among the isolates, indicating overall conserved chr 1 s for the bacterium. Nevertheless, numerous low similarity regions were detected, particularly throughout chr 1 in CHN108B. The regions in the largest LCB (2406 kb) added up to 160 kb in CHN108B, of which

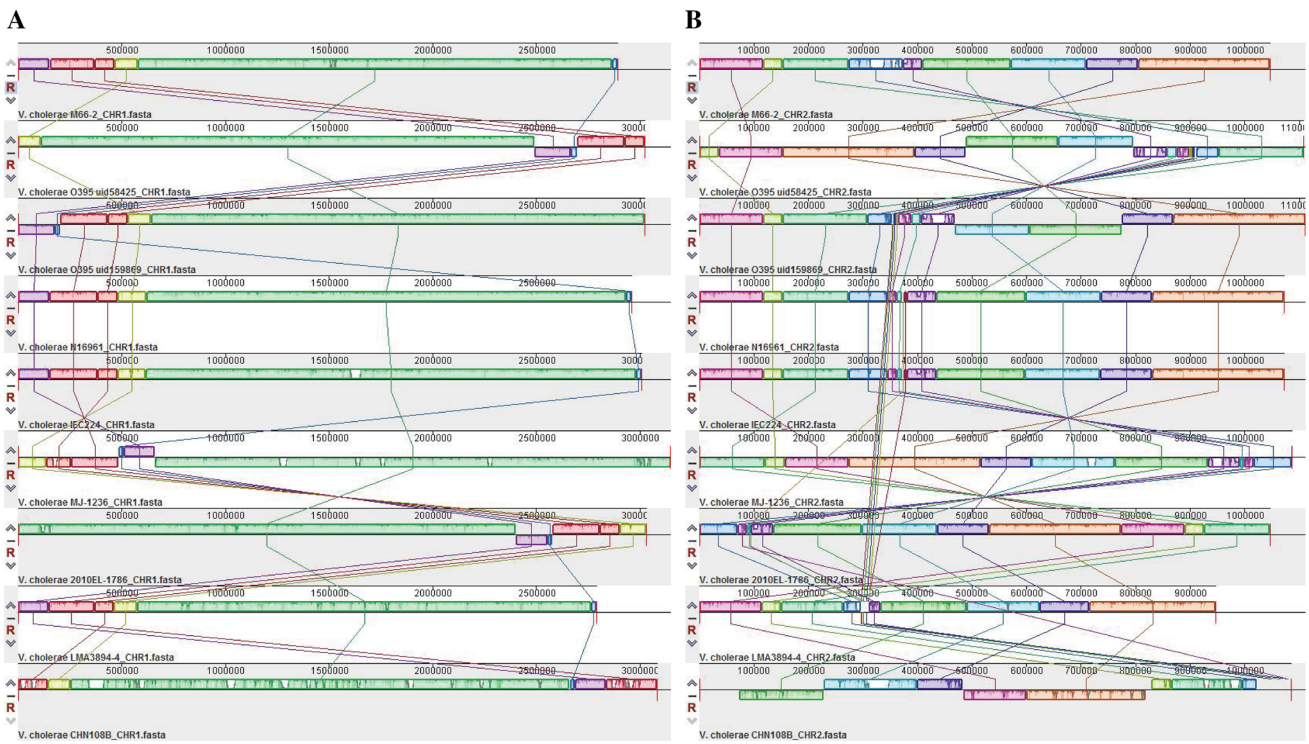

Fig. 2 The global genome sequence alignment of *V. cholerae* CHN108B and the other eight *V. cholerae* isolates. The complete genome sequence of *V. cholerae* CHN108B was obtained in this study, while those of the other eight isolates were retrieved from GenBank database with the accession numbers given in Table S1. The blocks in the same colour connected with lines represent colinear regions between the genomes, and the blocks in white represent low similar sequences between the genomes. **a** chr 1 s, **b** chr 2 s

the majority of the genes encoded hypothetical proteins. Interestingly, chr 1 from the LMA3984-4 isolate shared a similar LCB structure with the 1937 Makassar outbreak isolate M66-2 and the 7th pandemic isolate N16961; no chromosome rearrangement was found among their chr 1 s. Likewise, eleven LCBs were identified in the chr 2 s from the *V. cholerae* isolates (Fig. 2b). In marked contrast to their chr 1 s, chromosomal rearrangements appeared to have frequently occurred in the chr 2 s, as the number of LCBs with changed genomic positions was high and their length long. Some small LCBs were lost from some isolates, while some regions were identified as strain-specific and were not included within an LCB, but likely acquired by LGT. For example, a larger such region (~ 45 kb) was found in the chr 2 from CHN108B (chr 2: 301,618 bp to 346,735 bp), in which most of the genes were involved in secondary metabolism, transport, and catabolism, which may allow for the strain-specific metabolic capabilities.

The global genome sequence alignment of CHN108B against each of the other eight *V. cholerae* isolates was also performed using the MUMmer 3.2.3 software (data not shown). The resulting data were consistent with those generated by the Mauve 2.3.1 software. Our results indicate that the genetic information carried in chr 1 s was more conserved than that in chr 2 s from the *V. cholerae* isolates, which explained that the vast majority of genes for essential cell functions and pathogenicity are located on the large chromosome (Heidelberg et al. 2000) (see below).

Subsequently, Blastcluster analysis further revealed 2,409 conserved core genes shared among the nine *V. cholerae* genomes analysed in this study, 10.8% of which encoded proteins of unknown function. The functional classification was performed for the identified genes against the COGs database, and the results are illustrated in Supplementary Fig. S1. The conserved core genes fell into 24 cell functional categories, with the largest three fractions of genes being grouped into the function unknown (10.8%); amino acid transport and metabolism (9.3%); and signal transduction mechanisms (7.6%). Likewise, the accessory genes were also analysed (Fig. S1), and the interesting finding was that accessory genes classified as DNA replication, recombination and repair were 1.8-fold higher than conserved core genes, which may be suggestive of strain-specific genes for this cell function. Because some pathogenic *V. cholerae*-specific genes (O1/O139 serogroup) have been reported (Dziejman et al. 2002; Chun et al. 2009; Kim et al. 2015), in this study, we further analysed the strain-specific genes. Remarkably, the CHN108B genome had 110 such genes, the highest among the *V. cholerae* genomes analysed in this study. Further classification analysis revealed that 21.8% of the strain-specific genes in CHN108B were function unknown and general function prediction only, with the remaining genes presumably related to strain-specific metabolism (see above).

MGEs and genome plasticity

ISs

ISs are the shortest autonomously mobile elements (< 2.5 kb) and have a simple genetic organization that can insert at multiple sites in a target molecule (Mahillon and Chandler 1998). Various ISs were identified in the nine *V. cholerae* genomes analysed in this study (Supplementary Table S3). In the CHN108B genome, only a single 1.3-kb IS1358 element was identified in chr 2. This element was also found to exist as a single copy in the other *V. cholerae* genomes, but was located in chr 1. The *V. cholerae* isolates (except CHN108B) also carried nine types of ISs located in either chr 1 and/or chr 2. For example, the 6th cholera pandemic isolates O395 uid58425 and O395 uid159869 had the maximum number (19) of ISs. Among these, the ISVch1, ISVch5, ISVch4 and IS1004 homologs existed as multiple copies (2–7) in their host genomes. The latter two were also identified in multiple positions in the genomes of five other *V. cholerae* isolates, including 2010 EL-1786, IEC224, M66-2, MJ1236, N16961 and LMA3984-4. Our results suggested that these ISs were likely to be active and capable of jumping in the *V. cholerae* genomes. Additionally, the 2010 Haiti cholera outbreak isolate 2010 EL-1786 had the maximum types of ISs (7), though four of which existed as a single copy, including IS1358, ISVch1, ISShfr9 and ISVsa3. The latter two were absent from the other eight *V. cholerae* genomes.

Prophages

Phages are closely related to bacterial pathogenicity especially via the transfer of virulence factors (Heidelberg et al. 2000). Consistent with previous studies (Heidelberg et al. 2000; Feng et al. 2008; Chun et al. 2009), prophage gene clusters were identified in seven *V. cholerae* genomes analysed in this study but were absent from the LMA3984-4 and M66-2 genomes. One such cluster was found to be specific to CHN108B (Fig. 3). This prophage encompassed 29.0 kb (chr 1: 1,020,243–1,049,280 bp) and encoded 25 ORFs, including 16 coding for phage-related proteins and 8 for hypothetical proteins. It only had low partial sequence similarity with a 38.2-kb phage *Vibrio* VP882 found in a pandemic, *Vibrio parahaemolyticus* O3:K6 strain isolated in Osaka, Japan (Lan et al. 2009). To our knowledge, this study was the first to reveal this prophage in *V. cholerae*.

Previous studies have indicated that the major virulence factors from pathogenic *V. cholerae* strains (O1/O139 serogroup) are encoded by the 6.9-kb phage CTXΦ integrated

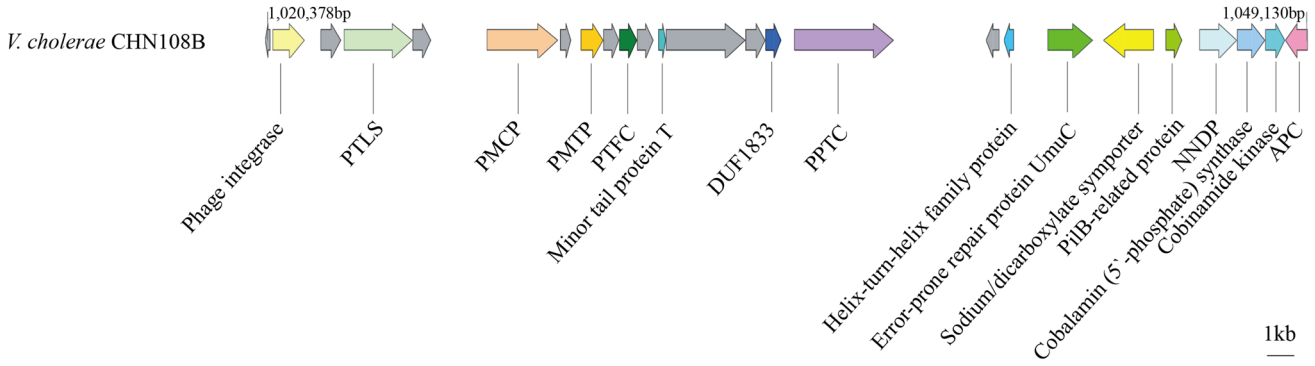

Fig. 3 The gene organizations of the prophage gene cluster identified in the *V. cholerae* CHN108B genome. The genes that encode hypothetical protein were shown in grey. *NNDP* nicotinate-nucleotide-dimethylbenzimidazole phosphoribosyltransferase, *PTLS* phage terminase large subunit (GpA), *PMCP* phage major capsid protein, HK97, *PMTP* prophage minor tail protein Z (GPZ), *PTFC* putative tail fibre component V of prophage, *DUF1833* domain of unknown function DUF1833, *PPTC* phage-related protein tail component, *APC* alpha-ribazole-5′-phosphate phosphatase CobC

in *V. cholerae* genomes (Rubin et al. 1998). In this study, the CTXΦ prophage was identified in six pathogenic *V. cholerae* isolates, including 2010 EL-1786, IEC224, MJ1236, N16961, O395 uid58425, and O395 uid159869. The latter two had two copies of the CTXΦ gene cluster. Although the identified CTXΦ-related sequences were different in size (6.3 to 58.0 kb) in the six *V. cholerae* genomes, they all contained genes encoding CT (*ctxAB*), accessory cholera enterotoxin (*ace*), zona occludens toxin (*zot*), and RS2 proteins responsible for phage replication (*rstA*), integration (*rstB*) and the regulation of site-specific recombination (*rstR*) (Waldor and Mekalanos 1996). Our results provided further evidence for the conserved minimum core virulence genes transferred by CTXΦ (Rubin et al. 1998). Interestingly, although the 7th cholera pandemic isolates N16961 and ICE224 were isolated from different geographical regions in the world: the former in Bangladesh in 1971 and the latter in Brazil in 1994, both strains contained almost identical CTXΦ sequences (30.5 kb, 95% sequence identity). Moreover, the 6th cholera pandemic isolates O395 uid58425 and O395 uid159869 had two copies of the CTX gene clusters; nevertheless, each copy showed a truncated CTXΦ molecular profile and was located in different host chrs. The sequences of these two copies in each isolate exhibited high nucleotide identity (> 90%) to the CTXΦ sequences identified in N16961 and IEC224. These results suggested that these strains may acquire cholera toxin genes from a common evolutionary ancestor via LGT, with subsequent genomic rearrangement into differential chrs in O395 uid58425 and O395 uid159869. Moreover, the identified CTXΦ sequence was longest (58.0 kb) in the 2010 Haiti cholera outbreak isolate 2010 EL-1786, which contained many ORFs encoding transposases and hypothetical proteins, suggesting extensive genetic recombination in this strain.

Another seven prophage gene clusters were identified in the genomes of the six clinical *V. cholerae* isolates, including 2010 EL-1786, IEC224, MJ1236, N16961, O395 uid58425 and O395 uid159869. For example, the O395 uid159869 genome had the maximum type of prophages showing sequence similarity with the *Aeromonas* phage 31, *Escherichia* phage D108 (2 copies), *Pseudomonas* phage B3, *Vibrio* phage CTX (2 copies), and *Vibrio* phage kappa. The MJ1236 and 2010 EL-1786 isolates also contained a *Vibrio* phage VvAW1 homologue, which infects *Vibrio vulnificus* (Nigro et al. 2012). Taken together, these data suggested extensive phage transmission between/among *Vibrio* species (*V. cholerae, V. parahaemolyticus*, and *V. vulnificus*), and bacterial genera (*Aeromonas, Escherichia*, and *Pseudomonas* and *Vibrio*).

ICEs

Self-transmissible ICEs and related elements can constitute a large proportion of bacterial chromosomes and bestow a wide range of phenotypes upon their hosts with carried gene cassettes (Wozniak and Waldor 2010). The SXT/R391 family of ICEs, one of the largest and most diverse set of ICEs, has been identified in Vibrionaceae isolates of clinical and environmental origins (Burrus et al. 2006; Wozniak and Waldor 2010). In this study, SXT/R391-like ICEs (73 to 108 kb) were identified in the genomes of three *V. cholerae* isolates CHN108B, 2010 EL-1786, and MJ1236, and designated ICE*Vch*CHN108B (chr 1: 338,087–411,356 bp), ICE*Vch*2010 EL-1786 (CHR1: 83,542–182,958 bp), and ICE*Vch*MJ1236 (chr 1:3,066,043–2,958,350 bp), respectively (Fig. 4). All the ICEs were located on the chr 1 of their hosts, and possessed most of the 52 predicted core genes for the highly conserved module structures that encode mating machinery for conjugation, intricate regulatory systems

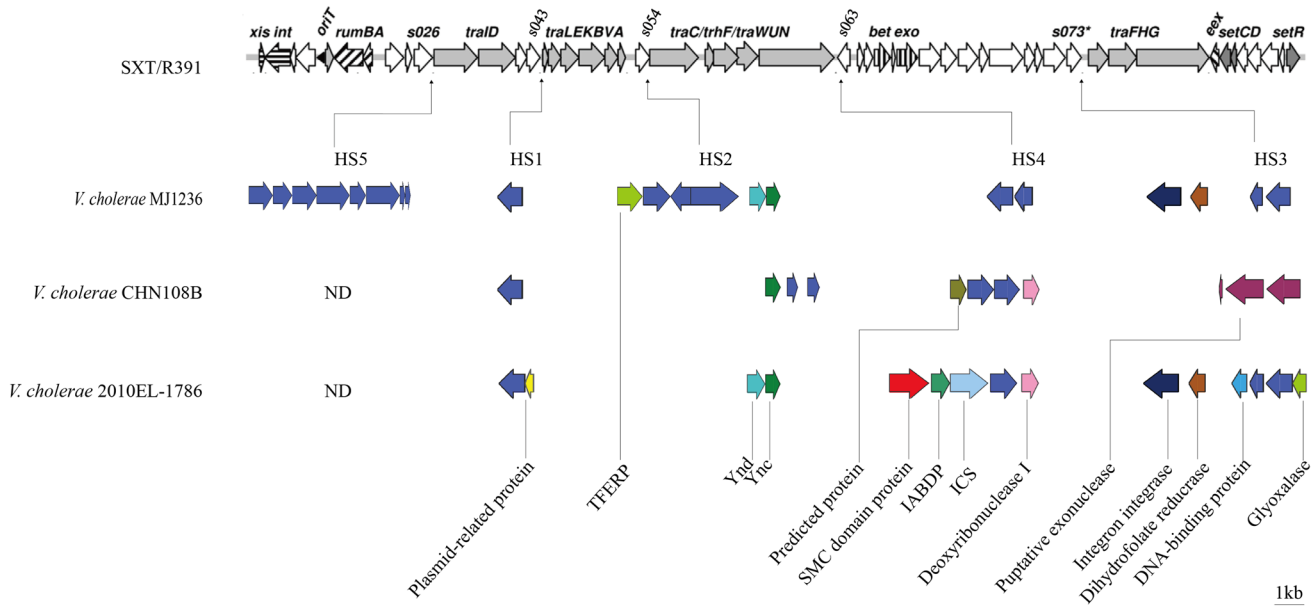

Fig. 4 Comparison of the accessory gene organizations in the ICEs identified in this study with the other known SXT/R391 ICEs. The gene organization of SXT/R391 ICEs was depicted by Wozniak et al. (2009). The genes that were inferred to encode homologous proteins were shown in the same colours in each hotspot region, and that encode hypothetical proteins were in blue. *TFERP* type 4 fimbriae expression regulatory protein, *IABDP* IstB ATP-binding domain-containing protein, *ICS* integrase catalytic subunit, *DNA-binding protein* DNA-binding HTH domain protein, *ND* not detected

to control their excision from the chromosome, as well as their self-transmissibility (Burrus and Waldor 2004). In the CHN108B genome, the identified ICE*Vch*CHN108B encompassed 73,269 bp and encoded 70 ORFs. Sequence analysis revealed that ICE*Vch*CHN108B had 96% nucleotide identify to ICE*Vch*Mex1 found in the *V. cholerae* isolate of an environmental origin (Burrus et al. 2006), which suggested a common ancestor shared by these two ICEs. No transposase-encoding genes were detected in the two ICEs, which may explain the relatively smaller size of ICE*Vch*CHN108B compared with the other identified ICEs in the *V. cholerae* genomes.

Accessory genes that are not required for transmission or other core ICE functions are restricted to insert into particular loci in several ICE families (Wozniak and Waldor 2010). In this study, we characterised DNA insertions in five hotspots (HS1 to HS5) and variable region III (VRIII) of the identified ICEs (Fig. 4), which are related to resistance determinants and other characterisation in previous reports (Wozniak and Waldor 2010). Consistent with previous research (Wozniak et al. 2009), the variable gene contents in the five hotspots were observed in the identified ICEs in this study (Fig. 4). For example, transposon-like structures carrying genes involved in trimethoprim resistance, DNA modification or recombination or repair in diverse putative restriction modification systems were found within the HS3 from SXT/R391-like ICEs (Wozniak et al. 2009). In this study, we found different HS3 molecular profiles between the ICEs of environmental and clinical origins. Three exonuclease-encoding genes were identified in the HS3 from ICE*Vch*CHN108B (*chr1_00339* and *chr1*_00440, *chr1*_00441), whereas no such gene was detected in ICE*Vch*MJ1236, and ICE*Vch*2010 EL-1786. These two ICEs of clinical origin shared similar gene contents in HS3, encoding an integrase, a dihydrofolate reductase type 1/trimethoprim resistance protein, and two hypothetical proteins. The HS3 from ICE*Vch*2010 EL-1786 also contained a DNA-binding HTH domain protein and a glyoxalase / bleomycin resistance protein. These results indicated that these ICEs likely conferred trimethoprim resistance to their clinical hosts MJ1236 and 2010 EL-1786. The latter also obtained resistance to bleomycin via the inserted genes in HS3 (Fig. 4).

SIs

In this study, all the *V. cholerae* genomes contained large (≥ 100 kb) SIs (Rowe-Magnus et al. 2002), except LMA3984-4. In the CHN108B genome, a 136-kb SI element (designated SI-CHN108B) was identified as located in chr 2 between two genes encoding a site-specific recombinase IntI4 (*chr2_00922*) and a transposase (*chr2_00095*), which was consistent with the other *V. cholerae* genomes. SI-CHN108B contained 182 predicted genes, 63% of which encoded hypothetical proteins. Higher percentages of hypothetical proteins (48–79%) were also observed in the other identified SIs in the *V. cholerae* genomes. It will be

interesting to determine their precise functions in the bacterium. Most of the remaining genes in SI-CHN108B were also detected in the other *V. cholerae* genomes.

Unlike the other identified SIs, the gene encoding the streptogramin A acetyltransferase VatD (*chr2_00933*) was found to be specific to SI-CHN108B. Streptogramin A is a polyunsaturated macrolactone compound that exhibits a moderate bacteriostatic activity by binding to the bacterial 50S ribosomal subunit and thereby blocking translation (Mast and Wohlleben 2014). Resistance to streptogramin A can be mediated by acetyltransferases (i.e., vatA to vatD) via acetylation of the only hydroxyl in its structure (Mast and Wohlleben 2014). In this study, our data indicated that SI-CHN108B conferred a streptogramin A resistance determinant to its host. Integrons can gain new cassettes at their 5′-end by incorporating at an *attI* site (Lan et al. 2008). Interestingly, in SI-CHN108B, the gene encoding the streptogramin A acetyltransferase was only 7 kb away from the *attI* site, implying that a more recent gene transfer event occurred in this SI element. This may explain why this resistance gene was absent from the other identified SIs in the *V. cholerae* genomes. Water environments contaminated with industrial pollutants may enhance selection for antimicrobial resistance and vice versa (Hu et al. 2015a). Our data may be suggestive of the inappropriate release of industrial wastes into the aquatic ecosystem where CHN108B survived.

Additionally, several toxin-associated genes were absent from SI-CHN108B, but present in some of the other SIs of clinical origin, including genes encoding a microcin immunity protein MccF, a toxin resistance protein, a haemagglutinin associated protein (HAP), a prevent-host-death protein (PHDP), and a glyoxalase/bleomycine resistance protein.

Pathogenic strain-specific genes and putative virulence genes

The search for virulence-associated genes in the nine *V. cholerae* genomes revealed distinct molecular profiles between the clinical and environmental isolates (Supplementary Table S4). We found the CT genes (*ctxAB*) and the *tcpIPHABQCRDSTEFJ* gene cluster in all the clinical isolates, except M66-2 lacked the CT genes. The *tcp* cluster is required for critical intestine colonisation factor TCP (Mohammadi-Barzelighi et al. 2011). These toxin genes were absent from CHN108B and LMA3984-4, which are of environmental origin. Although the presence of the CT- and TCP-encoding genes was the most obvious pathogenic strain-specific trait, some other toxin genes were also identified in most clinical isolates but absent from the environmental isolates. These included the *ace, zot*, and *acfABCD* gene cluster that is required for accessory colonisation factor (ACF).

Unexpectedly, some toxin genes were detected in most *V. cholerae* genomes of clinical or environmental origin (Table S4). These included the *rtxABCD* gene cluster for actin crosslinking repeats in toxin (RTX), which encodes an RTX toxin (*rtxA*), and its activator (*rtxC*) and transporters (*rtxBD*) (Lin et al. 1999); *hlyA*, encodes a secreted haemolysin A (HlyA), which has vacuolating activity (Figueroa-Arredondo et al. 2001) and *hap* encodes a haemagglutinin protease (Hap), which has direct effects on the proteins involved in maintaining the integrity of epithelial cell tight junctions (Wu et al. 2000). The gene clusters responsible for the biogenesis (*MshHIJKLMNEGF*) and structure (*MshBACD*) of mannose-sensitive haemagglutinin (MSHA), a type IV pilus, were also found to be common in all the *V. cholerae* isolates, which is inconsistent with a previous report showing that MSHA was unique to the El Tor biotype of *V. cholerae* (Heidelberg et al. 2000). MSHA is the receptor for a widespread filamentous bacteriophage in *V. cholerae* O139, which transfers a number of virulence genes (Jouravleva et al. 1998). Additionally, putative virulence-associated genes were also detected in the *V. cholerae* genomes of both clinical and environmental origin. They encoded proteins for the anti-phagocytosis (*cpsABCD*); chemotaxis (*cheW1B2A2ZY3*) (Hyakutake et al. 2005); colonisation (*pilABCD*, Type IV-A pilus); exoenzymes (*nanH*), flagellar biosynthesis, structure, motility and regulation (*flgABCDEFGHIJKLMN, fliADEFGHIJKLMNOPQRS, flhABF* and *motABXY, respectively*); and iron uptake (*hutR* and *vctA*) (Table S4). These genes could be candidate targets for the development of novel diagnostics, vaccines, and therapeutics for human cholera disease.

Substantial differences in antibiotic resistance genes

Numerous antimicrobial resistance genes were detected in the nine *V. cholerae* genomes analysed in this study (Supplementary Table S5). All the genomes contained the genes for the multidrug efflux pump AcrAB-TolC (Li et al. 2015), the cation-bound multidrug and toxic compound extrusion (MATE) transporter NorM (He et al. 2010), penicillin-binding proteins (Pbp1A/B) and beta-lactamase. The genes for resistance to ciprofloxacin (*pbp5*) and sulphonamide (*folP*) were also identified in all the *V. cholerae* genomes. Compared with the *V. cholerae* isolates of environmental origin, we found that the clinical isolates possessed more antibiotic resistance genes. Furthermore, the resistance gene patterns greatly differed among the 2010 EL-1786, MJ1236 and IEC224 clinical isolates. These clinical strains were isolated in more recent years (1994 to 2010) and have captured more resistance than those isolated earlier (1937 to 1994). For example, the genes for resistance to chloramphenicol (*cat*), sulphadoxine-pyrimethamine (*dhfrIII*), streptomycin

(*strAB*), tetracycline (*tetA*), and trimethoprim (*dfrA1*) were absent from the CHN108B and LMA3984-4 environmental isolates, but present in most of the three clinical isolates, indicating the increased development of antibiotic-resistant pathogenic *V. cholerae*, which likely arose from a hospital environment. Additionally, no resistance genes for other antimicrobial agents were identified in all the *V. cholerae* genomes, such as gentamicin, erythromycin, florfenicol, and vancomycin.

Differential secretion systems

Virulence factors are secreted by bacterial secretion systems (Wooldridge 2009). Six types of secretion systems (T1SS to T6SS) have been found in Gram-negative bacteria to date (Ehsani et al. 2009). Among these, T1SS is relatively simple and consists of only two or three proteins and directly translocates proteins (such as HlyA, RtxA, and lipases S-layer proteins) from the cytoplasm to the cell surface (Wooldridge 2009). The CHN108B genome had the genes (*rtxBDE, tolC*) required for T1SS, which is consistent with the other *V. cholerae* genomes. T2SSs were reported to maintain fitness in different ecological niches in *V. cholerae* and are involved in the extracellular export of CT and other proteins (Sikora 2013). The *gap* gene cluster for T2SS was identified in the clinical isolates, whereas the four T2SS genes (*gspCHIJ*) were missing in CHN108B. The results suggested that the degenerative T2SS probably resulted from the absence or loss of the TC and TCP in CHN108B and vice versa. T3SS is composed of approximately 30 different proteins to form a bacterial injectisome that spans the entire cell envelope, enabling the bacteria to inject bacterial effector proteins directly into the host cell cytoplasm (Coburn et al. 2007). Shin et al. reported that T3SS was essential for the rapidly fatal diarrhoeal disease caused by non-O1 and non-O139 *V. cholerae*, such as AM-19226 (Shin et al. 2011). However, in this study, T3SS genes were not detected in CHN108B (non-O1/O139 serogroup) or in the O1 serogroup isolates (except the *vcsS2* gene). T6SS are encoded by a cluster of 15–20 genes that is present with at least one copy in approximately 25% of all sequenced Gram-negative bacteria (Bingle et al. 2008; Basler et al. 2012). It has also been detected in *V. cholerae* and other pathogenic bacteria (Pukatzki et al. 2006). Previous studies have indicated that the genes *hcp* and *vgrG*, which encode a haemolysin co-regulated protein and a valine-glycine repeat, respectively, are essential for T6SS-dependent secretion in *V. cholerae* (Pukatzki et al. 2006). In CHN108B, these genes were identified in multiple copies, including four copies in chr 1 and a single *hcp* and three copies of *vgrG* in chr 2. Consistent with the other *V. cholerae* genomes, some genes encoding core components of the virulence-associated nanomachine T6SS were present in CHN108B, including IcmF, DotU, putative lipoprotein Lip, ClpV, and AAA^{+}ATPase. However, the IcmH-like protein and the regulatory proteins PpkA, Fhal and PppA were missing in CHN108B. The PppA-encoding gene was detected in all the clinical isolates, whereas the PpkA gene was missing in LMA3984-4 as well. Additionally, the genes encoding T4SS and T5SS proteins were absent from the *V. cholerae* genomes analysed in this study.

V. cholerae genomes shaped by the Yangtze River Estuary environment

To gain an insight into the impact of the Yangtze River Estuary environment upon the *V. cholerae* genomes, nine other *V. cholerae* isolates (CHN001f to 009f) recovered from the same environment as CHN108B were subjected for Illumina sequencing with a genome sequencing depth of 100-fold. The nine draft genome sequences were obtained with genome sizes in the range from 3.961 to 4.010 Mb. Comparison of accessory genes among the eighteen *V. cholerae* genomes analysed in this study (ten of the Yangzi River Estuary origin) revealed 28 estuary environment-specific genes. Of these, approximately half genes encoded hypothetical proteins, and the other half encoded enzymes (aminopeptidase, choline dehydrogenase and low molecular weitht phosphatase), transporters, antiporters, membrane protein and transcriptional regulators, which are probably related to the unique environment adaptation of *V. cholerae*.

Discussion

In the present study, we determined the whole genome sequence of the non-O1/O139 *V. cholerae* strain CHN108B isolated from surface water from the Yangtze River Estuary in China. Comparative genomics between CHN108B and the other eight *V. cholerae* isolates within serogroup O1 revealed interesting findings in the bacterial genome structure. CHN108B has the highest number of strain-specific genes when compared with the other *V. cholerae* isolates of clinical and environmental origin. Previous studies have indicated that the intercellular transmissibility of the MGEs with carried gene cassettes could constitute important driving forces in *Vibrios* genome evolution and speciation and mediate the emergence, resurgence and spread of multiple drug-resistant pathogens (Rowe-Magnus et al. 2002; Burrus and Waldor 2004; Siguier et al. 2014). In this study, various MGEs including ISs, prophages, ICEs, and SIs were identified in the *V. cholerae* genomes of clinical and environmental origin, indicating the bacterium undergoes extensive genetic recombination via LGT. For example, large MGEs carrying many genes of unknown function were detected in CHN108B, which may contribute to its unique genome traits.

The global emergence of multidrug-resistant Gram-negative bacteria is a growing threat to antibiotic therapy (Li et al. 2015). Consistent with previous reports (Li et al. 2015), numerous genes that mediate intrinsic and acquired multidrug resistance were identified in the *V. cholerae* genomes analysed in this study. The comparison of these genes revealed different antimicrobial resistance gene patterns between the environmental and clinical isolates, as well as among the clinical isolates originating from different cholera outbreak years. Our data provide further evidence for the increased development of resistance by pathogenic *V. cholerae*, which threatens the therapeutic options versus human cholera disease. The resistance was mediated by chromosomally encoded genes and the mechanisms potentially involve MGEs in clinical cases (Hochhut et al. 2001). Additionally, a relatively narrow resistance gene pattern yielded by LMA3984-4 was observed when compared with CHN108B, as some genes for resistance to fluoroquinolone (*gyrAB*), rifampin (*rpoB*), penicillin-binding protein (*pbp1A*), and streptogramin (*VatD*) were absent from the LMA3984-4 genome, but present in CHN108B, likely resulting from the significant difference in their geographic origin and selective pressure. High incidences of antibiotic resistant *V. cholerae* isolates of the Yangtze River Estuary origin have been reported previously and presumably arose from the abuse of drugs and the inappropriate release of industrial wastes into the environment (Song et al. 2013; He et al. 2015a).

Previous studies have indicated that accessory virulence factors whose functions in human cholera disease remain elusive may play an essential role in the fitness of *V. cholerae* (Faruque and Mekalanos 2012). Adhesion and colonisation in the human small intestine are also crucial for the pathogenic cycle of *V. cholerae*. The *cpsABCD* gene cluster responsible for capsular biosynthesis was identified in all the *V. cholerae* genomes analysed in this study and encodes a capsular polysaccharide biosynthesis glycosyltransferase (*cpsA*), a mannose-1-phosphate guanylyltransferase (*cpsB*), a polysaccharide export-like protein (*cpsC*), and an exopolysaccharide biosynthesis protein (*cpsD*). Additionally, the genes in the *acfABCD* gene cluster were all missing from the *V. cholerae* isolates of environmental origin. Moreover, the *pilA* gene from the *pilABCD* gene cluster was also missing from CHN108B and LMA3984-4. This gene cluster encodes proteins for the Type IV-A pilus, which is essential for the secretion of CT and Hap and colonisation of infant mice or adherence to HEp-2 cells (Fullner and Mekalanos 1999). Genes for the acquisition of nutrient iron also play a critical role in aiding the pathogenic bacteria in establishing and maintaining infection in hosts (Ratledge and Dover 2000). *V. cholerae* has multiple iron transport systems, one of which involves haem uptake through the outer membrane receptor HutA. The bacterium also encodes two additional TonB-dependent haem receptors, HutR and HasR (Mey and Payne 2001). In this study, unlike the clinical isolates, the *hasR* and *HutA* genes were absent from CHN108B, and the latter was also missing from LMA3984-4, which was consistent with the different niches between the clinical and environmental isolates. Furthermore, the gene *vctA* encoding an enterobactin receptor protein in another iron acquisition system (Mey et al. 2002) was identified in all the *V. cholerae* genomes. Taken together, our data revealed distinct virulence gene patterns between the *V. cholerae* isolates of clinical and environmental origin. Consistent with a previous report (Ceccarelli et al. 2015), our data also provided evidence for the presence of multiple virulence-associated genes in the non-O1/non-O139 *V. cholerae* isolates, such as *hlyA, rtxABCD, hap*, and *nanH* that encodes a neuraminidase. In the future research, it will be interesting to further address their functions in *V. cholerae* virulence.

Overall, this study is the first to describe the complete genome sequence of the non-O1/O139 *V. cholerae* CHN108B ($ct^{-}tcp^{-}$) isolated from the Yangtze River Estuary in China. It has 110 strain-specific genes, 21.8% of which encode hypothetical proteins, and the majority of the remaining encoded proteins are required for secondary metabolism and cell envelope and outer membrane biogenesis. Comparative genomics revealed numerous MGEs, including ISs, Prophages, ICEs, and SIs in the nine whole-genome *V. cholerae* isolates of clinical and environmental origin, indicating that the bacterium underwent extensive genetic recombination via LGT during its evolution. Moreover, our data revealed that some virulence-associated genes are present in most clinical isolates, but absent from the environmental isolates, such as *acfABCD, hasR, HutA* and *T6SS*, which could be candidate target genes for the development of novel diagnostics, vaccines, and therapeutics against human cholera disease. Additionally, comparative genomics provided the evidence for distinct antibiotic resistance gene patterns among the *V. cholerae* isolates. It should be noted that there are relatively wide resistance gene patterns yielded by the clinical isolates. Moreover, the clinical isolates (MJ1236 and 2010 EL-1786) isolated in more recent years have the highest number of antibiotic resistance genes responsible for resistance to multiple antimicrobial agents, indicating the ongoing and rising development of resistance among pathogenic *V. cholerae*, which is a real threat to public health. Therefore, future studies on more *V. cholerae* genomes of environmental origin will enhance our understanding of the epidemiology and pathogenesis of this severe water-borne pathogen worldwide.

Author contributions PY, HZ, WH, YD, XP, QS, LX, LC participated in the design and or discussion of the study. LG performed genome gap closure, gene annotation and comparative genomic analysis. HZ directed the sequencing and genome analysis. WG and WH assisted in the genome sequencing and gene annotation. LG, PY, YT, YW, LC

analysed the data. LG, LC wrote the manuscript. All authors read and approved the final manuscript for publication.

Funding This work was supported by a grant No.B-9500-10-0004 from Shanghai Municipal Education Commission, China, a grant No.17050502200 from Science and Technology Commission of Shanghai Municipality, China, and grants No.31271830 and No.31671946 from National Natural Science Foundation of China. We thank YL and HF for the help in DNA preparation, and WY for the environment-specific gene analysis in this study.

Data availability The genome sequence data were submitted to GenBank under the BioSample Accession Number SUB2559601, 4497099, 4497317, 4497328, 4497348, 4497357, 4497359, 4497361, 4497365 and 4497394.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

Banerjee R, Das B, Balakrish Nair G, Basak S (2014) Dynamics in genome evolution of *Vibrio cholerae*. Infect Genet Evol 23:32–41

Basler M, Pilhofer M, Henderson GP, Jensen GJ, Mekalanos JJ (2012) Type VI secretion requires a dynamic contractile phage tail-like structure. Nature 483:182–186

Bingle LE, Bailey CM, Pallen MJ (2008) Type VI secretion: a beginner's guide. Microbiol 11:3–8

Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF (2011) Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. MBio 2:e00335–e00310

Burrus V, Waldor MK (2004) Shaping bacterial genomes with integrative and conjugative elements. Res Microbiol 155:376–386

Burrus V, Quezada-Calvillo R, Marrero J, Waldor MK (2006) SXT-related integrating conjugative element in New world *Vibrio cholerae*. Appl Environ Microbiol 72:3054–3057

Ceccarelli D, Chen A, Hasan NA, Rashed SM, Huq A, Colwell RR (2015) Non-O1/non-O139 *Vibrio cholerae* carrying multiple virulence factors and *V. cholerae* O1 in the Chesapeake Bay, Maryland. Appl Environ Microbiol 81:1909–1918

Chatterjee S, Ghosh K, Raychoudhuri A, Chowdhury G, Bhattacharya MK, Mukhopadhyay AK, Ramamurthy T, Bhattacharya SK, Klose KE, Nandy RK (2009) Incidence,virulence factors, and clonality among clinical strains of non-O1, non-O139 *Vibrio cholerae* isolates from hospitalized diarrheal patients in Kolkata, India. J Clin Microbiol 47:1087–1095

Chen L, Brügger K, Skovgaard M, Redder P, She Q, Torarinsson E et al (2005) The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota. J Bacteriol 87:4992–4999

Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley et al (2009) Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. Proc Natl Acad Sci USA 106:15442–15447

Coburn B, Sekirov I, Finlay BB (2007) Type III secretion systems and disease. J Clin Microbiol 20:535–549

Colwell RR (1996) Global climate and infectious disease: the cholera paradigm. Science 274:2025–2031

Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–4641

Dutilh BE, Thompson CC, Vicente AC, Marin MA, Lee C, Silva GG, Schmieder R, Andrade BG, Chimetto L, Cuevas D, Garza DR, Okeke IN, Aboderin AO, Spangler J, Ross T, Dinsdale EA, Thompson FL, Harkins TT, Edwards RA (2014) Comparative genomics of 274 *Vibrio cholerae* genomes reveals mobile functions structuring three niche dimensions. BMC Genom 15:654

Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, Mekalanos JJ (2002) Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. Proc Natl Acad Sci USA 99:1556–1561

Ehsani S, Rodrigues CD, Enninga J (2009) Turning on the spotlight: using light to monitor and characterize bacterial effector secretion and translocation. Curr Opin Microbiol 12:24–30

Faruque SM, Mekalanos JJ (2012) Phage-bacterial interactions in the evolution of toxigenic *Vibrio cholerae*. Virulence 3:556–565

Feng L, Reeves PR, Lan R, Ren Y, Gao C, Zhou Z, Ren Y, Cheng J, Wang W, Wang J, Qian W, Li D, Wang L (2008) A recalibrated molecular clock and independent origins for the cholera pandemic clones. PLoS One 3:e4053

Figueroa-Arredondo P, Heuser JE, Akopyants NS, Morisaki JH, Giono-Cerezo S, Enríquez-Rincón F, Berg DE (2001) Cell vacuolation caused by *Vibrio cholerae* hemolysin. Infect Immun 69:1613–1624

Fullner KJ, Mekalanos JJ (1999) Genetic characterization of a new type IV-A pilus gene cluster found in both classical and El Tor biotypes of *Vibrio cholerae*. Infect Immun 67:1393–1404

Garrine M, Mandomando I, Vubil D, Nhampossa T, Acacio S, Li S, Paulson JN, Almeida M, Domman D, Thomson NR, Alonso P, Stine OC (2017) Minimal genetic change in *Vibrio cholerae* in Mozambique over time: multilocus variable number tandem repeat analysis and whole genome sequencing. PLoS Negl Trop Dis 11(6):e0005671

Garza DR, Thompson CC, Loureiro EC, Dutilh BE, Inada DT, Junior EC, Cardoso JF, Nunes MR, de Lima CP, Silvestre RV, Nunes KN, Santos EC, Edwards RA, Vicente AC (2012) Genome-wide study of the defective sucrose fermenter strain of *Vibrio cholerae* from the Latin Americancholera epidemic. PloS ONE 7:e37283

He X, Szewczyk P, Karyakin A, Evin M, Hong WX, Zhang Q, Chang G (2010) Structure of a cation-bound multidrug and toxic compound extrusion transporter. Nature 467:991–994

He Y, Tang Y, Sun F, Chen L (2015a) Detection and characterization of integrative and conjugative elements (ICEs)-positive *Vibrio cholerae* isolates from aquacultured shrimp and the environment in Shanghai, China. Mar Pollut Bull 101:526–532

He Y, Wang H, Chen L (2015b) Comparative secretomics reveals novel virulence-associated factors of *Vibrio parahaemolyticus*. Front Microbiol 6:707

Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ et al (2000) DNA sequence of both chromosomes of the cholera pathogen. *Vibrio cholerae* Nature 406:477–483

Hochhut B, Lotfi Y, Mazel D, Faruque SM, Woodgate R, Waldor MK (2001) Molecular analysis of antibiotic resistance gene clusters in

Vibrio cholerae O139 and O1 SXT constins. Antimicrob Agents Chem 45:2991–3000

Hyakutake A, Homma M, Austin MJ, Boin MA, Ha¨se CC, Kawagishi I (2005) Only one of the five CheY homologs in *Vibrio cholerae* directly switches flagellar rotation. J Bacteriol 187:8403–8410

Imamura D, Morita M, Sekizuka T, Mizuno T, Takemura T, Yamashiro T, Chowdhury G, Pazhani GP, Mukhopadhyay AK, Ramamurthy T, Miyoshi SI, Kuroda M, Shinoda S, Ohnishi M (2017) Comparative genome analysis of VSP-II and SNPs reveals heterogenic variation in contemporary strains of *Vibrio cholerae* O1 isolated from cholera patients in Kolkata, India. PLoS Negl Trop Dis 11(2):e0005386

Jouravleva EA, McDonald GA, Marsh JW (1998) The *Vibrio cholerae* mannose-sensitive hemagglutinin is the receptor for a filamentous bacteriophage from *V. cholerae* O139. Infect Immun 66:2535–2539

Kaper JB, Morris JG Jr, Levine MM (1995) Cholera. Clin Michrobiol Rev 8:48–86

Kim EJ, Lee CH, Nair GB, Kim DW (2015) Whole-genome sequence comparisons reveal the evolution of *Vibrio cholerae* O1. Trends Microbiol 23:479–489

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C et al (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12

Lan SF, Huang CH, Chang CH, Liao WC, Lin IH, Jian WN, Wu YG, Chen SY, Wong HC (2009) Characterization of a new plasmid-like prophage in a pandemic *Vibrio parahaemolyticus* O3:K6 strain. Appl Environ Microbiol 75:2659–2667

Li X-Z, Plésiat P, Nikaido H (2015) The challenge of efflux-mediated antibiotic resistance in Gram-negative bacteria. Clin Microbiol Rev 28:337–418

Lin W, Fullner KJ, Clayton R, Sexton JA, Rogers MB, Calia KE, Calderwood SB, Fraser C, Mekalanos JJ (1999) Identification of a *Vibrio cholerae* RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. Proc Natl Acad Sci USA 96:1071–1076

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Mahillon J, Chandler M (1998) Insertion sequences. Microbiol Mol Biol Rev 62:725–774

Mast Y, Wohlleben W (2014) Streptogramins—two are better than one. Int J Med Microbiol 304:44–50

Mazel D, Dychinco B, Webb VA, Davies J (1998) A distinctive class of integron in the *Vibrio cholerae* genome. Science 280:605–608

Metzker M (2005) Emerging technologies in DNA sequencing. Genome Res 15:1767–1776

Mey AR, Payne SM (2001) Haem utilization in *Vibrio cholerae* involves multiple TonB-dependent haem receptors. Mol Microbiol 42:835–849

Mey AR, Wyckoff EE, Oglesby AG, Rab E, Taylor RK, Payne SM (2002) Identification of the *Vibrio cholerae* enterobactin receptors VctA and IrgA: IrgAis not required for virulence. Infect Immun 70:3419–3426

Mohammadi-Barzelighi H, Bakhshi B, Lari AR, Pourshafie MR (2011) Characterization of pathogenicity island prophage in clinical and environmental strains of *Vibrio cholerae*. J Med Microbiol 60:1742–1749

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JL, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462–465

Nigro OD, Culley AI, Steward GF (2012) Complete genome sequence of bacteriophage VvAW1, which infects *Vibrio vulnificus*. Stand Genomic Sci 6:415–426

Okada K, Na-Ubol M, Natakuathung W, Roobthaisong A, Maruyama F, Nakagawa I, Chantaroj S, Hamada S (2014) Comparative genomic characterization of a Thailand-Myanmar isolate, MS6, of *Vibrio cholerae* O1 El Tor, which is phylogenetically related to a "US Gulf Coast" clone. PLoS One 9(6):e98120

Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, Heidelberg JF, Mekalanos JJ (2006) Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. Proc Natl Acad Sci USA 103:1528–1533

Ratledge C, Dover LG (2000) Iron metabolism in pathogenic bacteria. Annu Rev Microbiol 54:881–941

Reimer AR, Van Domselaar G, Stroika S, Walker M, Kent H, Tarr C, Talkington D, Rowe L, Olsen-Rasmussen M, Frace M, Sammons S, Dahourou GA, Boncy J, Smith AM, Mabon P, Petkau A, Graham M, Gilmour MW, Gerner-Smidt P (2011) Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. Emerg Infect Dis 17:2113–2121

Rowe-Magnus DA, Guerout AM, Mazel D (2002) Super-integrons. Mol Microbiol 43:1657–1669

Rowe-Magnus DA, Guerout AM, Biskri L, Bouige P, Mazel D (2003) Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. Genome Res 13:428–442

Rubin EJ, Lin W, Mekalanos JJ, Waldor MK (1998) Replication and integration of a *Vibrio cholerae* cryptic plasmid linked to the CTX prophage. Mol Microbiol 28:1247–1254

Sá LL, Vale ER, Garza DR, Vicente AC (2012) *Vibrio cholerae* O1 from superficial water of the Tucunduba Stream, Brazilian Amazon. Braz J Microbiol 43:635–638

Shin OS, Tam VC, Suzuki M, Ritchie JM, Bronson RT, Waldor MK, Mekalanos JJ (2011) Type III secretion is essential for the rapidly fatal diarrheal disease caused by non-O1, non-O139 *Vibrio cholerae*. MBio 2:e00106–e00111

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34:D32–D36

Siguier P, Gourbeyre E, Chandler M (2014) Bacterial Insertion Sequences: Their genomic impact and diversity. FEMS Microbiol Rev 38:865–891

Sikora AE (2013) Proteins secreted via the type II secretion system: smart strategies of *Vibrio cholerae* to maintain fitness in different ecological niches. PLoS Pathog 9:e1003126

Song Y, Yu P, Li B, Pan Y, Zhang X, Cong J, Zhao Y, Wang H, Chen L (2013) The mosaic accessory gene structures of the SXT/R391-like integrative conjugative elements derived from *Vibrio* spp. isolated from aquatic products and environment in the Yangtze River Estuary, China. BMC Microbiol 12:214

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28

Thompson FL, Iida T, Swings J (2004) Biodiversity of Vibrios. Microbiol Mol Biol Rev 68:403–431

Wachsmuth IK, Blake P, Olsvik O (eds) (1994) *Vibrio cholerae* and cholera: Molecular to global perspective. ASM Press

Waldor MK, Mekalanos JJ (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 272:1910–1914

Waldor MK, Tschape H, Mekalanos JJ (1996) A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in *Vibrio cholerae* O139. J Bacteriol 178:4157–4165

Wooldridge K (2009) Bacterial secreted proteins: secretory mechanisms and role in pathogenesis. Caister Academic Press. ISBN 978-1-904455-42-4

Wozniak RA, Waldor MK (2010) Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol 8:552–563

Wozniak RA, Fouts DE, Spagnoletti M, Colombo MM, Ceccarelli D, Garriss G, De'ry C, Burrus V, Waldor MK (2009) Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs. PLoS Genet 5:e10007865

Wu Z, Nybom P, Magnusson K-E (2000) Distinct effects of *Vibrio cholerae* haemagglutinin/protease on the structure and localization of the tight junction-associated proteins occludin and ZO-1. Cell Microbiol 2:11–17

Yi Y, Lu N, Liu F, Li J, Zhang R, Jia L, Jing H, Xia H, Yang Y, Zhu B, Hu Y, Cui Y (2014) Genome sequence and comparative analysis of a *Vibrio cholerae* O139 strain E306 isolated from a cholera case in China. Gut Pathog 6(1):3

Zhang T, Gou Q (2014) Traveling wave solutions for epidemic cholera model with disease-related death. Sci World J 2014:409730

Affiliations

Li Gong[1] · Pan Yu[1] · Huajun Zheng[2] · Wenyi Gu[2] · Wei He[3] · Yadong Tang[1] · Yaping Wang[1] · Yue Dong[4] · Xu Peng[5] · Qunxin She[5] · Lu Xie[6] · Lanming Chen[1]

Li Gong
emo_jun@163.com

Pan Yu
p-yu@shou.edu.cn

Huajun Zheng
zhenghj@chgc.sh.cn

Wenyi Gu
guwy@chgc.sh.cn

Wei He
shchgc@gmail.com

Yadong Tang
tangyy0402@126.com

Yaping Wang
1327656510@qq.com

Yue Dong
Yue.Dong-1@ou.edu

Xu Peng
peng@bio.ku.dk

Qunxin She
qunxin@bio.ku.dk

[1] Key Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage and Preservation (Shanghai), China Ministry of Agriculture, College of Food Science and Technology, Shanghai Ocean University, Shanghai, People's Republic of China

[2] Shanghai-MOST Key Laboratory of Disease and Health Genomics, Chinese National Human Genome Center at Shanghai, Shanghai, People's Republic of China

[3] Shanghai Hanyu Bio-lab, Shanghai, People's Republic of China

[4] University of Oklahoma, Norman, USA

[5] Department of Biology, University of Copenhagen, Copenhagen, Denmark

[6] Shanghai Center for Bioinformation Technology, Shanghai, People's Republic of China